#4

35.C14035 PATENT APPLICATION

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:	)	
	:	Examiner: NYA
TAKAFUMI MIZUNO	)	
	:	Group Art Unit: 2776
Application No.: 09/449,706	)	
	:	
Filed: November 24, 1999	)	
	:	
For: DOCUMENT TYPE DEFINITION GENERATING METHOD AND APPARATUS, AND STORAGE MEDIUM FOR STORING PROGRAM	)	February 4, 2000

Assistant Commissioner for Patents
Washington, D.C. 20231

CLAIM TO PRIORITY

Sir:

Applicant hereby claims priority under the International Convention and all rights to which he is entitled under 35 U.S.C. § 119 based upon the following Japanese Priority Application:

10-336378 filed November 26, 1998

A certified copy of the priority document is enclosed.

Applicant's undersigned attorney may be reached in our New York office by telephone at (212) 218-2100. All correspondence should continue to be directed to our address given below.

Respectfully submitted,

Jack M. Arnold

Attorney for Applicant

Registration No. 25,823

FITZPATRICK, CELLA, HARPER & SCINTO
30 Rockefeller Plaza
New York, New York 10112-3801
Facsimile: (212) 218-2200

NY_MAIN 59171 v 1

CFo 14035 (EP) US.
/m

09/449,706

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: １９９８年１１月２６日

出 願 番 号
Application Number: 平成１０年特許願第３３６２７８号

出 願 人
Applicant (s): キヤノン株式会社

１９９９年１２月１７日

特許庁長官
Commissioner,
Patent Office 近藤隆彦

【書類名】　特許願

【整理番号】　3798113

【提出日】　平成10年11月26日

【あて先】　特許庁長官　殿

【国際特許分類】　G06F 17/27

【発明の名称】　文書型定義生成方法および文書処理装置並びに文書型定義生成用プログラムを記録した記録媒体

【請求項の数】　24

【発明者】

【住所又は居所】　東京都大田区下丸子３丁目３０番２号　キヤノン株式会社内

【氏名】　水野　貴史

【特許出願人】

【識別番号】　000001007

【氏名又は名称】　キヤノン株式会社

【代理人】

【識別番号】　100077481

【弁理士】

【氏名又は名称】　谷　義一

【選任した代理人】

【識別番号】　100088915

【弁理士】

【氏名又は名称】　阿部　和夫

【手数料の表示】

【予納台帳番号】　013424

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　　要約書　1

【包括委任状番号】　9703598

【プルーフの要否】　　要

【書類名】　明細書

【発明の名称】　文書型定義生成方法および文書処理装置並びに文書型定義生成用プログラムを記録した記録媒体

【特許請求の範囲】

【請求項１】　複数の電子化文書から構造化文書に変換された文書、あるいは元々構造化文書であった文書から、その文書構造を表現するために該文書中に埋め込まれたタグのおかれている前後の行（１行以上）の字下げ、空白行という物理的構造を判断する物理的構造判断ステップと、

前記タグの意味的構造を判断する意味的構造判断ステップとを有し、

前記物理的構造判断と前記意味的構造判断の結果に元づいて構造化文書の文書型定義を生成することを特徴とする文書型定義生成方法。

【請求項２】　前記タグの物理的構造の判断ステップにおいて、タグの位置関係を認識することを特徴とする請求項１に記載の文書型定義生成方法。

【請求項３】　前記タグの位置関係の認識は、字下げを行うことによって引用を表す引用中に行われる字下げや１行あるいは複数行を常に飛ばしている場合のような物理的構造検出にとって無意味な構造を取り除く処理も含むことを特徴とする請求項２に記載の文書型定義生成方法。

【請求項４】　前記タグの意味的構造の判断ステップにおいて、意味情報データベースを参照して文書中の語句の繋がりや単語の種類等を元に当該タグの意味的構造を検出することを特徴とする請求項１ないし３のいずれかに記載の文書型定義生成方法。

【請求項５】　前記物理的構造判断と前記意味的構造判断の結果に元づいて、異なった名称のタグ間において物理的構造と意味的構造の一致度合いについての類似度を求め、該類似度の値が所定のしきい値以上の場合は、それらタグは同一のタグと見なして最終的に文書型定義を生成するリストから一方のタグを削除する処理を行う冗長性排除ステップを有することを特徴とする請求項１ないし４のいずれかに記載の文書型定義生成方法。

【請求項６】　前記冗長性排除ステップは、前記類似度の値が所定のしきい値未満の場合に、同一名称を持つ異なる意味のタグの名称変更を行うことを特徴

とする請求項５に記載の文書型定義生成方法。

【請求項７】　前記冗長性排除ステップは、タグ同士の相対的な位置関係やタグの包含関係についての物理的な情報や、あるいはタグが表わす意味による意味的な情報を前記類似度の尺度として用いることを特徴とする請求項６記載の文書型定義生成方法。

【請求項８】　同一名称の開始タグと終了タグ間に存在する文の語句解析を行い、当該タグ中に含まれるべき情報を得て、該情報を元に文書型定義の生成を行う文書型定義生成ステップを有することを特徴とする請求項１ないし７のいずれかに記載の文書型定義生成方法。

【請求項９】　複数の電子化文書から構造化文書に変換された文書、あるいは元々構造化文書であった文書から、その文書構造を表現するために該文書中に埋め込まれたタグのおかれている前後の行（１行以上）の字下げ、空白行という物理的構造を判断する物理的構造判断手段と、

前記タグの意味的構造を判断する意味的構造判断手段とを有し、

前記物理的構造判断と前記意味的構造判断の結果に元づいて構造化文書の文書型定義を生成することを特徴とする文書処理装置。

【請求項１０】　前記タグの物理的構造の判断手段は、タグの位置関係を認識することを特徴とする請求項９に記載の文書処理装置。

【請求項１１】　前記タグの位置関係の認識は、字下げを行うことによって引用を表す引用中に行われる字下げや１行あるいは複数行を常に飛ばしている場合のような物理的構造検出にとって無意味な構造を取り除く処理も含むことを特徴とする請求項１０に記載の文書処理装置。

【請求項１２】　前記タグの意味的構造の判断手段は、意味情報データベースを参照して文書中の語句の繋がりや単語の種類等を元に当該タグの意味的構造を検出することを特徴とする請求項９ないし１１のいずれかに記載の文書処理装置。

【請求項１３】　前記物理的構造判断と前記意味的構造判断の結果に元づいて、異なった名称のタグ間において物理的構造と意味的構造の一致度合いについての類似度を求め、該類似度の値が所定のしきい値以上の場合は、それらタグは

同一のタグと見なして最終的に文書型定義を生成するリストから一方のタグを削除する処理を行う冗長性排除手段を有することを特徴とする請求項９ないし１２のいずれかに記載の文書処理装置。

【請求項１４】　前記冗長性排除手段は、前記類似度の値が所定のしきい値未満の場合に、同一名称を持つ異なる意味のタグの名称変更を行うことを特徴とする請求項１３に記載の文書処理装置。

【請求項１５】　前記冗長性排除手段は、タグ同士の相対的な位置関係やタグの包含関係についての物理的な情報や、あるいはタグが表わす意味による意味的な情報を前記類似度の尺度として用いることを特徴とする請求項１４記載の文書処理装置。

【請求項１６】　同一名称の開始タグと終了タグ間に存在する文の語句解析を行い、当該タグ中に含まれるべき情報を得て、該情報を元に文書型定義の生成を行う文書型定義生成手段を有することを特徴とする請求項９ないし１５のいずれかに記載の文書処理装置。

【請求項１７】　コンピュータによって実行される文書型定義生成用プログラムを記録した記録媒体であって、該プログラムはコンピュータに対し、

複数の電子化文書から構造化文書に変換された文書、あるいは元々構造化文書であった文書から、その文書構造を表現するために該文書中に埋め込まれたタグのおかれている前後の行（１行以上）の字下げ、空白行という物理的構造を判断させ、

前記タグの意味的構造を判断させ、

前記物理的構造判断と前記意味的構造判断の結果に元づいて構造化文書の文書型定義を生成させることを特徴とする文書型定義生成用プログラムを記録した記録媒体。

【請求項１８】　前記プログラムはコンピュータに対し、

前記タグの物理的構造の判断において、タグの位置関係を認識させることを特徴とする請求項１７記載の文書型定義生成用プログラムを記録した記録媒体。

【請求項１９】　前記プログラムはコンピュータに対し、

前記タグの位置関係の認識において、字下げを行うことによって引用を表す引

用中に行われる字下げや１行あるいは複数行を常に飛ばしている場合のような物理的構造検出にとって無意味な構造を取り除く処理行わせることを特徴とする請求項１８に記載の文書型定義生成用プログラムを記録した記録媒体。

【請求項２０】　前記プログラムはコンピュータに対し、

前記タグの意味的構造の判断において、意味情報データベースを参照して文書中の語句の繋がりや単語の種類等を元に当該タグの意味的構造を検出させることを特徴とする請求項１７ないし１９のいずれかに記載の文書型定義生成用プログラムを記録した記録媒体。

【請求項２１】　前記プログラムはコンピュータに対し、

前記物理的構造判断と前記意味的構造判断の結果に元づいて、異なった名称のタグ間において物理的構造と意味的構造の一致度合いについての類似度を求めさせ、該類似度の値が所定のしきい値以上の場合は、それらタグは同一のタグと見なして最終的に文書型定義を生成するリストから一方のタグを削除させることを特徴とする請求項１７ないし２０のいずれかに記載の文書型定義生成用プログラムを記録した記録媒体。

【請求項２２】　前記プログラムはコンピュータに対し、

前記類似度の値が所定のしきい値未満の場合に、同一名称を持つ異なる意味のタグの名称変更を行わせることを特徴とする請求項２１に記載の文書型定義生成用プログラムを記録した記録媒体。

【請求項２３】　前記プログラムはコンピュータに対し、

タグ同士の相対的な位置関係やタグの包含関係についての物理的な情報や、あるいはタグが表わす意味による意味的な情報を前記類似度の尺度として用いさせることを特徴とする請求項２２記載の文書型定義生成用プログラムを記録した記録媒体。

【請求項２４】　前記プログラムはコンピュータに対し、

同一名称の開始タグと終了タグ間に存在する文の語句解析を行わせ、当該タグ中に含まれるべき情報を得て、該情報を元に文書型定義の生成を行わせることを特徴とする請求項１７ないし２３のいずれかに記載の文書型定義生成用プログラムを記録した記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、パーソナルコンピュータやワードプロセッサ等で実行される電子化文書の処理技術に関し、特に構造化文書の文書型定義を生成する方法および文書処理装置並びに文書型定義生成用プログラムを記録した記録媒体に関する。

【０００２】

【従来の技術】

近年、パーソナルコンピュータやワードプロセッサ等で作成された電子化された文書が溢れかえっている。このような電子化文書に対して、統一的に扱え、文書に意味情報を付与する構造化文書の導入が進められている。しかしながら、この構造化文書を記述するには、予め決められた規則（文書型定義）に従わなければならないことが多い。

【０００３】

しかし、従来では、タグの置かれている物理的観点、意味的観点から文書型定義を生成する方法はなかった。

【０００４】

【発明が解決しようとする課題】

文書型定義を予め用意しておかなくても、構造化文書を記述できるが、文書の記述を行うユーザ各々が勝手にタグを使ってしまうと、タグに付与された意味情報も正しく扱うことができなくなってしまう。また、逆に同一の種類と考えられる文書に対して複数の文書型定義が用意されている場合には、タグに対する冗長性が存在してしまう可能性がある。

【０００５】

本発明の目的は、上記のような課題を解決すべく、タグに与えられた意味情報を正しく扱うことができ、タグに対する冗長性を排除した文書型定義を生成することが可能な文書型定義生成方法および文書処理装置並びに文書型定義生成用プログラムを記録した記録媒体を提供することにある。

【０００６】

【課題を解決するための手段】

上記目的を達成するため、請求項１の文書型定義生成方法の発明は、複数の電子化文書から構造化文書に変換された文書、あるいは元々構造化文書であった文書から、その文書構造を表現するために該文書中に埋め込まれたタグのおかれている前後の行（１行以上）の字下げ、空白行という物理的構造を判断する物理的構造判断ステップと、前記タグの意味的構造を判断する意味的構造判断ステップとを有し、前記物理的構造判断と前記意味的構造判断の結果に元づいて構造化文書の文書型定義を生成することを特徴とする。

【０００７】

ここで、好ましくは、前記タグの物理的構造の判断ステップにおいて、タグの位置関係を認識する。

【０００８】

また、好ましくは、前記タグの位置関係の認識は、字下げを行うことによって引用を表す引用中に行われる字下げや１行あるいは複数行を常に飛ばしている場合のような物理的構造検出にとって無意味な構造を取り除く処理も含む。

【０００９】

また、好ましくは、前記タグの意味的構造の判断ステップにおいて、意味情報データベースを参照して文書中の語句の繋がりや単語の種類等を元に当該タグの意味的構造を検出する。

【００１０】

また、好ましくは、前記物理的構造判断と前記意味的構造判断の結果に元づいて、異なった名称のタグ間において物理的構造と意味的構造の一致度合いについての類似度を求め、該類似度の値が所定のしきい値以上の場合は、それらタグは同一のタグと見なして最終的に文書型定義を生成するリストから一方のタグを削除する処理を行う冗長性排除ステップを有する。

【００１１】

また、好ましくは、前記冗長性排除ステップは、前記類似度の値が所定のしきい値未満の場合に、同一名称を持つ異なる意味のタグの名称変更を行う。

【００１２】

また、好ましくは、前記冗長性排除ステップは、タグ同士の相対的な位置関係やタグの包含関係についての物理的な情報や、あるいはタグが表わす意味による意味的な情報を前記類似度の尺度として用いる。

【００１３】

また、好ましくは、同一名称の開始タグと終了タグ間に存在する文の語句解析を行い、当該タグ中に含まれるべき情報を得て、該情報を元に文書型定義の生成を行う文書型定義生成ステップを有する。

【００１４】

上記目的を達成するため、請求項９の文書処理装置の発明は、複数の電子化文書から構造化文書に変換された文書、あるいは元々構造化文書であった文書から、その文書構造を表現するために該文書中に埋め込まれたタグのおかれている前後の行（１行以上）の字下げ、空白行という物理的構造を判断する物理的構造判断手段と、前記タグの意味的構造を判断する意味的構造判断手段とを有し、前記物理的構造判断と前記意味的構造判断の結果に元づいて構造化文書の文書型定義を生成することを特徴とする。

【００１５】

ここで、好ましくは、前記タグの物理的構造の判断手段は、タグの位置関係を認識する。

【００１６】

また、好ましくは、前記タグの位置関係の認識は、字下げを行うことによって引用を表す引用中に行われる字下げや１行あるいは複数行を常に飛ばしている場合のような物理的構造検出にとって無意味な構造を取り除く処理も含む。

【００１７】

また、好ましくは、前記タグの意味的構造の判断手段は、意味情報データベースを参照して文書中の語句の繋がりや単語の種類等を元に当該タグの意味的構造を検出する。

【００１８】

また、好ましくは、前記物理的構造判断と前記意味的構造判断の結果に元づい

て、異なった名称のタグ間において物理的構造と意味的構造の一致度合いについての類似度を求め、該類似度の値が所定のしきい値以上の場合は、それらタグは同一のタグと見なして最終的に文書型定義を生成するリストから一方のタグを削除する処理を行う冗長性排除手段を有する。

【００１９】

また、好ましくは、前記冗長性排除手段は、前記類似度の値が所定のしきい値未満の場合に、同一名称を持つ異なる意味のタグの名称変更を行う。

【００２０】

また、好ましくは、前記冗長性排除手段は、タグ同士の相対的な位置関係やタグの包含関係についての物理的な情報や、あるいはタグが表わす意味による意味的な情報を前記類似度の尺度として用いる。

【００２１】

また、好ましくは、同一名称の開始タグと終了タグ間に存在する文の語句解析を行い、当該タグ中に含まれるべき情報を得て、該情報を元に文書型定義の生成を行う文書型定義生成手段を有する。

【００２２】

上記目的を達成するため、請求項１７の記録媒体の発明は、コンピュータによって実行される文書型定義生成用プログラムを記録した記録媒体であって、該プログラムはコンピュータに対し、複数の電子化文書から構造化文書に変換された文書、あるいは元々構造化文書であった文書から、その文書構造を表現するために該文書中に埋め込まれたタグのおかれている前後の行（１行以上）の字下げ、空白行という物理的構造を判断させ、前記タグの意味的構造を判断させ、前記物理的構造判断と前記意味的構造判断の結果に元づいて構造化文書の文書型定義を生成させることを特徴とする。

【００２３】

また、好ましくは、前記プログラムはコンピュータに対し、前記タグの物理的構造の判断において、タグの位置関係を認識させる。

【００２４】

また、好ましくは、前記プログラムはコンピュータに対し、前記タグの位置関

係の認識において、字下げを行うことによって引用を表す引用中に行われる字下げや１行あるいは複数行を常に飛ばしている場合のような物理的構造検出にとって無意味な構造を取り除く処理行わせる。

【００２５】

また、好ましくは、前記プログラムはコンピュータに対し、前記タグの意味的構造の判断において、意味情報データベースを参照して文書中の語句の繋がりや単語の種類等を元に当該タグの意味的構造を検出させる。

【００２６】

また、好ましくは、前記プログラムはコンピュータに対し、前記物理的構造判断と前記意味的構造判断の結果に元づいて、異なった名称のタグ間において物理的構造と意味的構造の一致度合いについての類似度を求めさせ、該類似度の値が所定のしきい値以上の場合は、それらタグは同一のタグと見なして最終的に文書型定義を生成するリストから一方のタグを削除させる。

【００２７】

また、好ましくは、前記プログラムはコンピュータに対し、前記類似度の値が所定のしきい値未満の場合に、同一名称を持つ異なる意味のタグの名称変更を行わせる。

【００２８】

また、好ましくは、前記プログラムはコンピュータに対し、タグ同士の相対的な位置関係やタグの包含関係についての物理的な情報や、あるいはタグが表わす意味による意味的な情報を前記類似度の尺度として用いさせる。

【００２９】

また、好ましくは、前記プログラムはコンピュータに対し、同一名称の開始タグと終了タグ間に存在する文の語句解析を行わせ、当該タグ中に含まれるべき情報を得て、該情報を元に文書型定義の生成を行わせる。

【００３０】

本発明は、上記構成により、タグの置かれている前後の行の字下げ、空白行という物理的構造を判断し、タグの意味的構造を判断して、タグの位置関係を認識することにより文書の型定義を行うようにしているので、タグに与えられた意味

情報を正しく扱うことが可能となる。

【００３１】

更に、本発明では、タグに対する冗長性の排除、及び、同一名称で異なる意味のタグの名称変更を行った文書型定義を生成することが可能となる。

【００３２】

【発明の実施の形態】

以下、図面を参照して本発明の実施の形態を詳細に説明する。

【００３３】

（第１の実施の形態）

本発明に係る文書型定義生成方法を実現する一実施形態の処理手順を図１のフローチャートに示す。なお、以下に述べる手順は、プログラムコード形態でフロッピディスク（ＦＤ）、ハードディスク、光ディスク、光磁気ディスク、ＣＤ－ＲＯＭ等の記録媒体に記録されることができ、ＣＰＵで読み出されて実行されるものとする。また、この文書型定義生成方法は一般的なハード構成のパーソナルコンピュータやワードプロセッサ等で実行可能であるので、この方法を実行するハードウエアの具体的なシステム構成の説明は省略する。

【００３４】

まず、ステップ１０１において構造化文書の入力が行われる。ここで与える構造化文書の一例を図２の（ａ）に示す。この構造化文書の中から、次のステップ１０２において各々のタグ位置が検出され、最初に“＜Ｔｉｔｌｅ＞”から順番にタグ番号が振られる。

【００３５】

続いて、ステップ１０３において、上記文書における物理的構造の検出が行われる。例えば、図２の（ｂ）において、与えた段落を表すタグ“＜Ｐａｒａ＞”において模式的に表現されているように、字下げで始まる文章群を段落と見なすという特徴の検出を行う。このような物理的構造を検出するための処理手順一例を図３のフローチャートに示す。

【００３６】

ステップ３１で文書中の字下げが行われている行を見つけ出し、次のステップ

３２でその行に続く文章群を検出する。このときに字下げが行われている行から次の字下げが行われている行、あるいは空白行の直前の行までを文章群とすることができる。また、字下げを行うことによって引用を表す引用中に行われている字下げ（２重の字下げ）や、一行あるいはそれ以上に行を常に飛ばして記述されている場合には、ステップ３２の中で、文書全体のパターンから、物理的構造検出にとって無意味な構造を取り除いて処理を行う。

【００３７】

図１のステップ１０４では、入力された構造化文書が持つ意味的構造を検出する。一例として、上記の図２の（ａ）において、“＜Ｓｅｃｔｉｏｎ＞”タグは、先頭に「１．」，「２．」，「３．」といった形式をしている。ここで意味的に、“＜Ｓｅｃｔｉｏｎ＞”タグは、「数字．」からなると推察できる。この意味的構造を検出する処理手順の一例を図４のフローチャートに示す。

【００３８】

まず、ステップ４１において、文書中の全ての単語、記号について意味情報ＤＢ（データベース）４３と通信をして、文書中の語間の繋がりや単語・記号の種類を与える。この結果を元に、次のステップ４２において、各タグにおいて見られる意味的な構造を検出する。

【００３９】

次に、図１のステップ１０５において、処理すべきタグの開始を最初に出現したタグとし、ステップ１０６でタグの処理が全て終了したか否かを判断する。

【００４０】

タグの処理が全て終了していない場合は、ステップ１０７に移行して、現在処理しているタグと、上記ステップ１０３，ステップ１０４において検出された物理的構造・意味的構造の情報との統合を行う。ここで、統合とは、現在、処理しているタグに関わる行に物理的な、意味的な特徴が存在するならば、タグとその情報を結びつけることをいう。次のステップ１０８では、次に出現したタグに処理を移し、上記ステップ１０６に戻る。

【００４１】

ステップ１０６においてタグの処理が全て終了したと判断した場合は、ステッ

プ１０９に移行し、異なった名称のタグの間において類似度を求め、その類似度が予め設定したあるしきい値以上の場合にはそれらタグは互いに同一のタグと見なして、一方のタグを生成する文書型定義に出現しないようにする。この類似度を求めて同じ内容を持つタブであるか否かを判定する処理手順の一例を図５のフローチャートに示す。

【００４２】

まず、ステップ５１では、異なる名称をもつタグＡ，Ｂの類似度を算出する。この算出方法は、物理的構造が一致する場合は１、物理的構造が完全に一致しないが一部が一致する場合は、その一致した割合に対応する１未満の値にし、意味的構造についても同様の考えをあてはめ、その和を２で割ったものを類似度d_{AB}とする。

【００４３】

ステップ５１で求めた上記の類似度d_{AB}に対して、次のステップ５２において、予め与えたしきい値δとの比較を行う。その類似度d_{AB}がδ未満のときは、ステップ５４に飛んで、次の組み合せを試みる。

【００４４】

その類似度d_{AB}がδより大きいときは、ステップ５３に移行して、最終的に文書型定義を生成するリストからタグＢを削除し、冗長性の排除を行う。また、フローチャートとして示さないが、同一名称を持つ異なる意味のタグの名称変更も図１のステップ１０９中で行う。

【００４５】

ステップ５３の処理が終了したらステップ５４へ進み、全てのタグの組み合わせを試したか否かを判断し、全てのタグの組み合わせを試していない場合はステップ５１に戻り、上記と同様に、タグＡａ　，Ａｂ　間での類似度を求めて、その類似度の値d_{AaAb}がしきい値δ未満のときは、タグＡｂ　の名称の変更を行う。全てのタグの組み合わせを試した場合は本サブルーチンの処理を終了して図１のメインルーチンに戻る。

【００４６】

図１のステップ１１０では、同一名称の開始タグと終了タグ間の文の語句につ

いての解析を行い、タグ中に含まれるべき情報を得る。この解析結果を用いて、次のステップ１１１において、文書型定義の生成を行う。

【００４７】

（第２の実施の形態）

上述した本発明の第１の実施形態においては、文書中の物理の構造、意味的構造を本文（タグ以外の文）に対して行ってきたが、本発明はその限りではない。

【００４８】

例えば、タグ同士の相対的な位置関係やタグの包含関係についての物理的な情報や、あるいはタグが表わす意味による意味的な情報を利用して類似度の尺度として用いてもよい。

【００４９】

（ 他の実施形態）

なお、本発明は、複数の機器（例えば、ホストコンピュータ、インターフェース機器、リーダ、プリンタなど）から構成されるコンピュータシステムに適用しても、１つの機器からなる装置（例えば、ワードプロセッサ、複写機、ファクシミリ装置など）に適用してもよい。

【００５０】

また、本発明の目的は、前述した実施の形態の機能を実現するソフトウエアのプログラムコードを記録した記録媒体（記憶媒体）を、システムあるいは装置に供給し、そのシステムあるいは装置のコンピュータ（またはＣＰＵやＭＰＵ）が記録媒体に格納されたプログラムコードを読み出し、実行することによっても、達成されることは言うまでもない。

【００５１】

この場合、記録媒体から読み出されたプログラムコード自体が前述した実施の形態の機能を実現することになり、そのプログラムコードを記録した記録媒体は本発明を構成することになる。

【００５２】

そのプログラムコードを記録し、またテーブル等の変数データを記録する記録媒体としては、例えばフロッピディスク（ＦＤ）、ハードディスク、光ディスク

、光磁気ディスク、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、磁気テープ、不揮発性のメモリカード（ＩＣメモリカード）、ＲＯＭなどを用いことができる。

【００５３】

また、コンピュータが読み出したプログラムコードを実行することにより、前述の実施の形態の機能が実現されるだけでなく、そのプログラムコードの指示に基づいて、コンピュータ上で稼動しているＯＳ（オペレーティングシステム）などが実際の処理の一部または全部を行ない、その処理によって前述した実施の形態の機能が実現される場合も含まれることは言うまでもない。

【００５４】

【発明の効果】

以上説明したように、本発明によれば、タグの置かれている前後の行の字下げ、空白行という物理的構造を判断し、タグの意味的構造を判断して、タグの位置関係を認識することにより文書の型定義を行うようにしているので、タグに与えられた意味情報を正しく扱うことが可能となる。

【００５５】

更に、本発明によれば、タグに対する冗長性の排除、及び、同一名称で異なる意味のタグの名称変更を行った文書型定義を生成することが可能となる。

【図面の簡単な説明】

【図１】

本発明に係る文書型定義生成方法を実現する一実施形態の処理手順を示すフローチャートである。

【図２】

本システムに与える構造化文書データの一例を示す説明図である。

【図３】

図１のステップ１０３における物理的構造解析の処理手順の一例を示すフローチャートである。

【図４】

図１のステップ１０４における意味的構造解析の処理手順の一例を示すフローチャートである。

【図５】

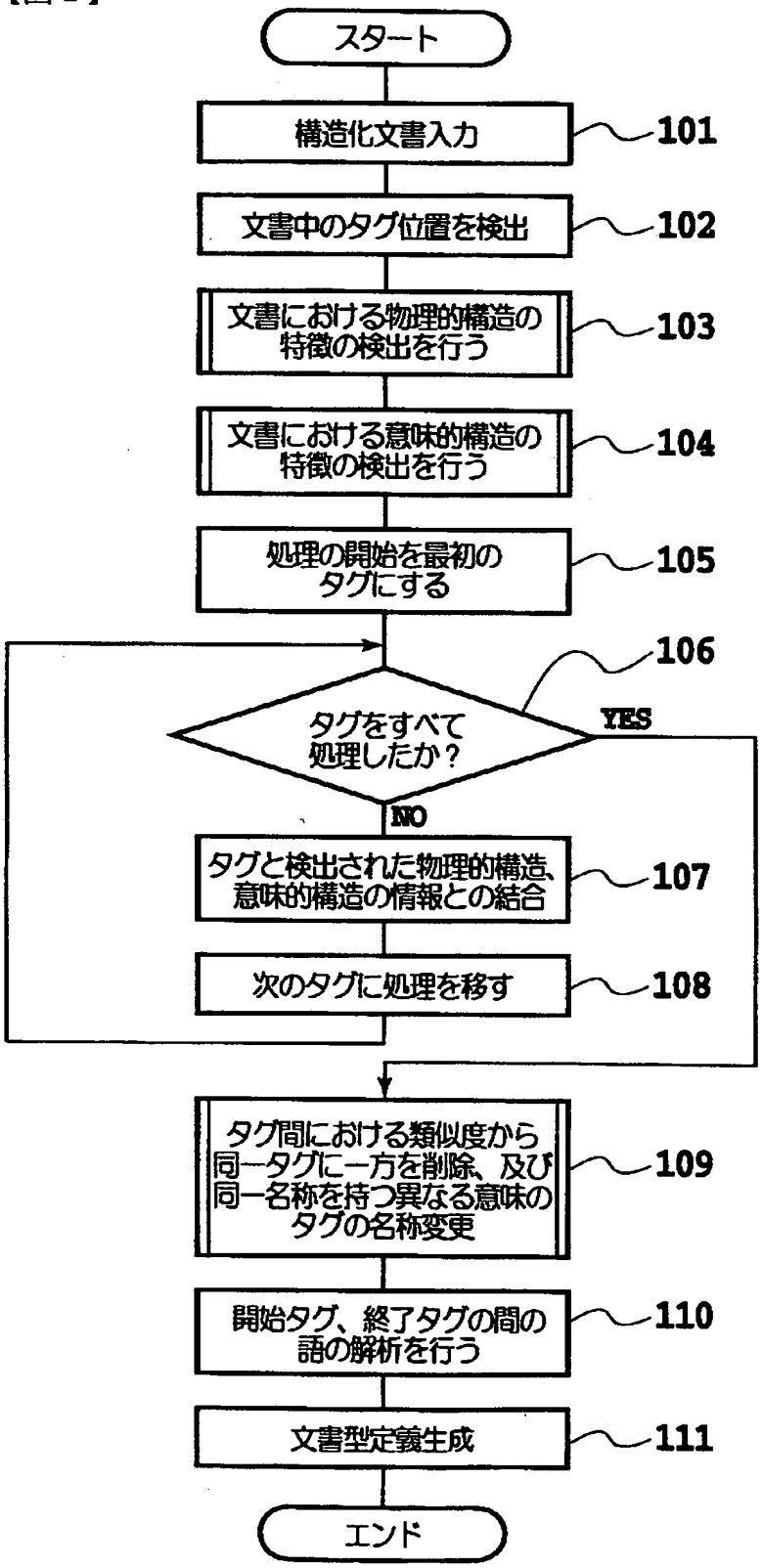
図１のステップ１０９におけるタグ間の類似度からタグの冗長性排除をする処理手順の一例を示すフローチャートである。

【符号の説明】

４３　意味情報ＤＢ（データベース）

【書類名】　　図面

【図１】


スタート
構造化文書入力
101
文書中のタグ位置を検出
102
文書における物理的構造の特徴の検出を行う
103
文書における意味的構造の特徴の検出を行う
104
処理の開始を最初のタグにする
105
106
タグをすべて処理したか？
YES
NO
タグと検出された物理的構造、意味的構造の情報との結合
107
次のタグに処理を移す
108
タグ間における類似度から同一タグに一方を削除、及び同一名称を持つ異なる意味のタグの名称変更
109
開始タグ、終了タグの間の語の解析を行う
110
文書型定義生成
111
エンド

【図２】

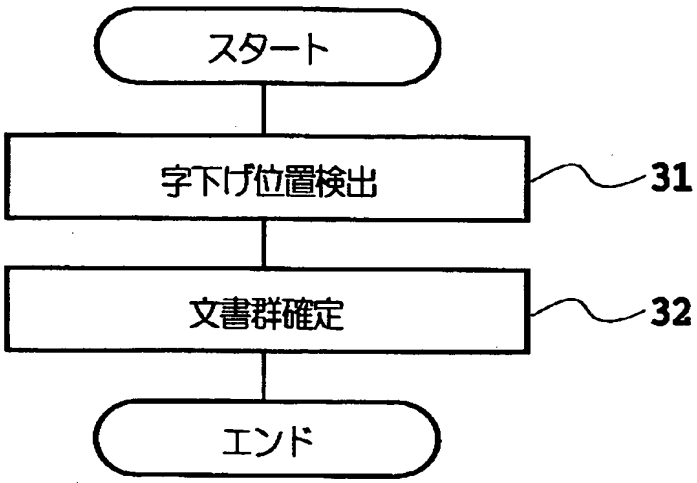
(a)

```
<Title>テレビの使用説明書</Title>
<Date> 1998.2.1</Date>
<Author>山田太郎</Author>
<Body>
    <Section>1.コンセントを差す</Section>
    <Section>2.電源を入れる </Section>
    <Section>3.チャネルを合わせる</Section>
</Body>
```

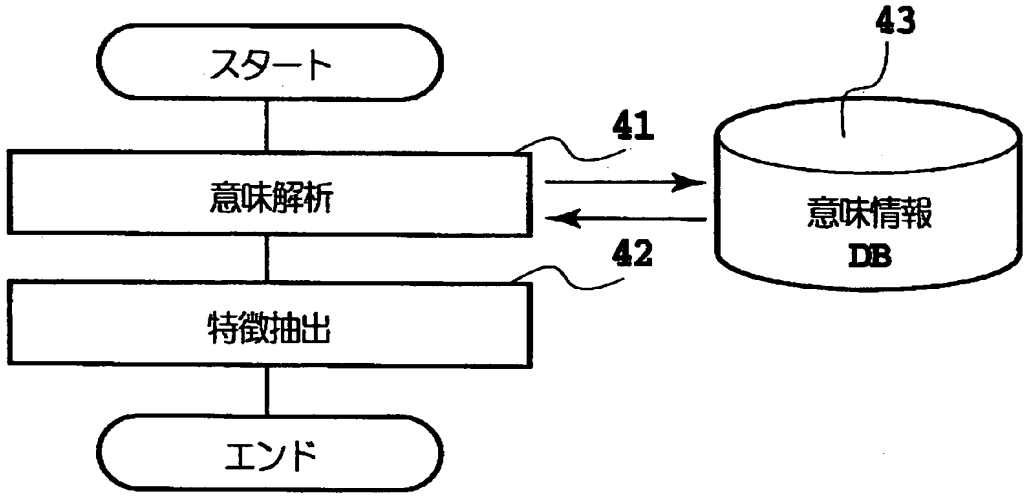
(b)

```
<Para>
    ―――――
    ―――――
    ―――――
</Para>
```

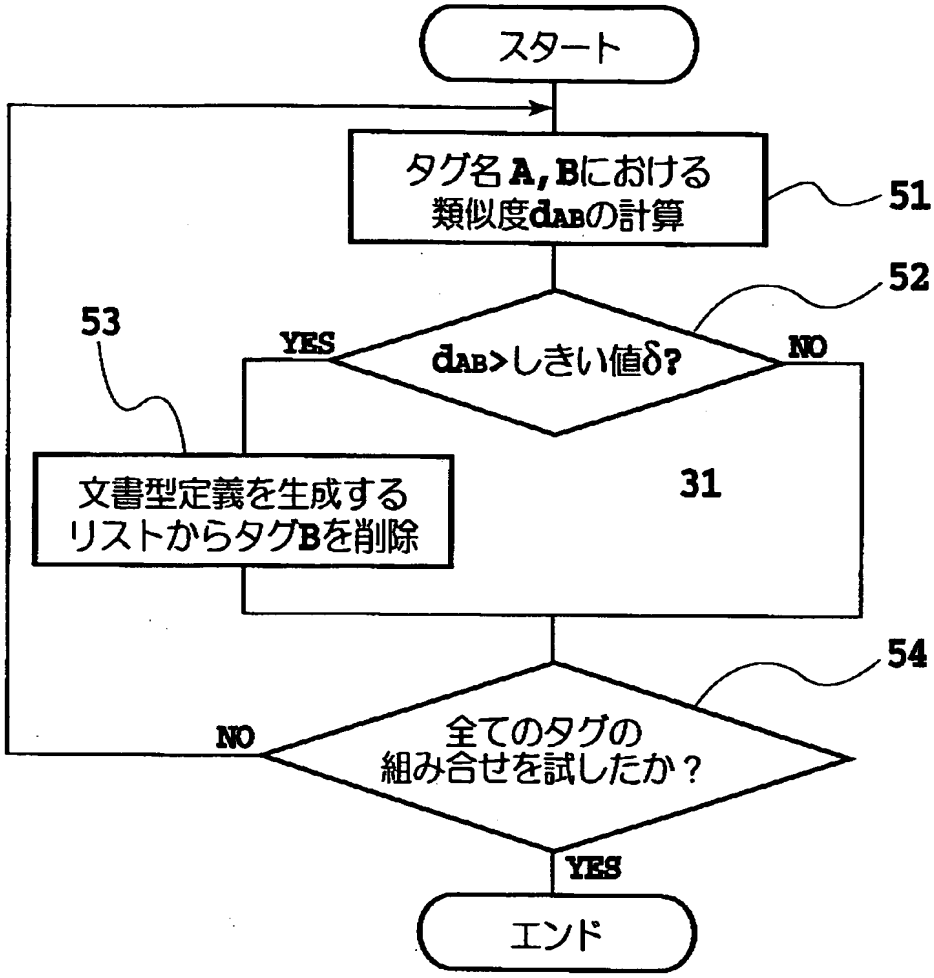
【図３】


スタート
字下げ位置検出
31
文書群確定
32
エンド

【図４】


スタート
意味解析
41
42
特徴抽出
エンド
43
意味情報
DB

【図５】



出証特平１１－３０８８５７９

【書類名】　　要約書

【要約】

【課題】　タグに与えられた意味情報を正しく扱うことができ、タグに対する冗長性を排除した文書型定義を生成する。

【解決手段】　構造化文書の文書構造を表現するために文書中に埋め込まれたタグのおかれている前後の行（１行以上）の字下げ、空白行という物理的構造を判断するステップ１０３と、タグの意味的構造を判断するステップ１０４と、異なった名称のタグ間において類似度を求め、類似度があるしきい値以上の場合は、同一のタグと見なして最終的に文書型定義を生成するリストから一方のタグを削除するステップ１０９と、同一名称の開始タグと終了タグ間に存在する文の語句解析を行い、タグ中に含まれるべき情報を得て、文書型定義の生成を行うステップ１１０、１１１とを有する。

【選択図】　　図１

【書類名】　職権訂正データ
【訂正書類】　特許願

＜認定情報・付加情報＞
【特許出願人】
【識別番号】　000001007
【住所又は居所】　東京都大田区下丸子３丁目３０番２号
【氏名又は名称】　キヤノン株式会社
【代理人】　申請人
【識別番号】　100077481
【住所又は居所】　東京都港区赤坂２丁目６番２０号　谷・阿部特許事務所
【氏名又は名称】　谷　義一
【選任した代理人】
【識別番号】　100088915
【住所又は居所】　東京都港区赤坂２丁目６番２０号　谷・阿部特許事務所
【氏名又は名称】　阿部　和夫

出　願　人　履　歴　情　報

識別番号　　〔０００００１００７〕

１．変更年月日　　１９９０年　８月３０日
［変更理由］　新規登録
住　所　東京都大田区下丸子３丁目３０番２号
氏　名　キヤノン株式会社